Lustre in CETA-CIEMAT

Implementation, configuration and integration

Author: Alfonso Pardo Diaz
Event: European Lustre Workshop
Lugar / Date: Paris, 27/09/2011

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

CENTRO EXTREMEÑO DE TECNOLOGÍAS AVANZADAS
CETA Ciemat

FEDER
Fondo Europeo de Desarrollo Regional
Una manera de hacer Europa

INDICE

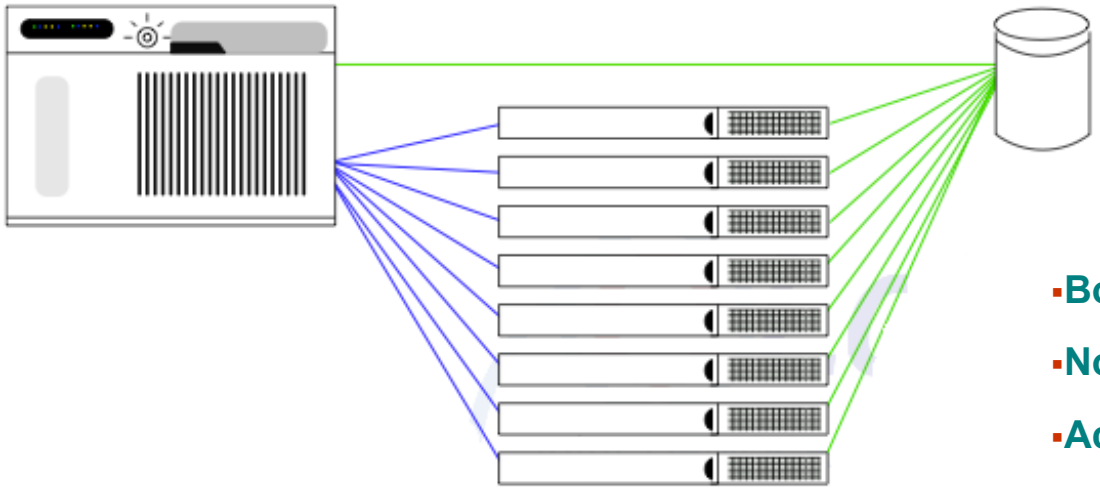
Lustre in CETA-CIEMAT

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
CENTRO EXTREMEÑO DE TECNOLOGÍAS AVANZADAS
CETA Ciemat
FEDER
Fondo Europeo de Desarrollo Regional
Una manera de hacer Europa

1 Motivation

ÍNDICE

- **Why?**
  - **A typical cluster is installed with a Master node, several computing nodes and shared storage.**



AND WE HAVE A LOT OF SERVERS OVER THE ETHERNET !!!

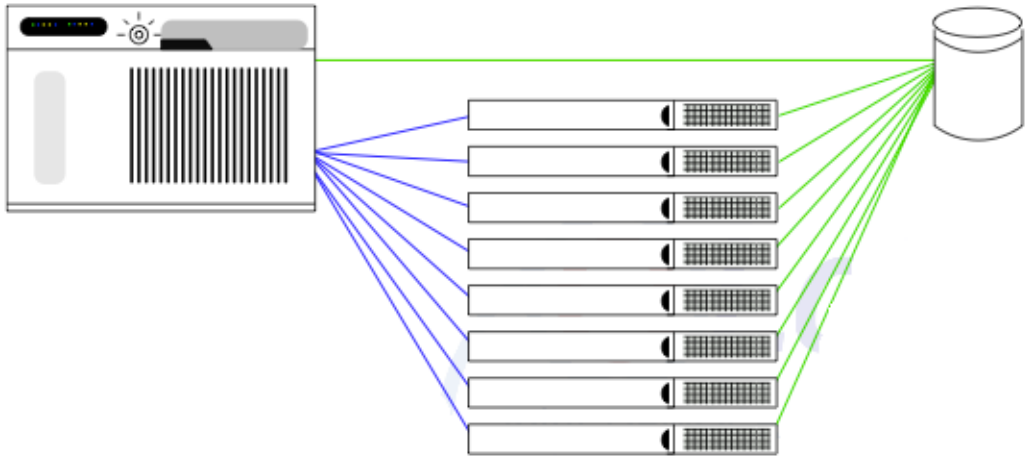
- **Bottleneck in storage server.**
- **No easy grow.**
- **Access data from client:**
- **NFS: bad for large-scale files number.**
- **SAN: HBA for clients and server: too expensive.**
- **Concurrence access files: locks.**

1 Motivation

ÍNDICE

- **Previously: NFS: bad for large-scale files number.**
  - Direct access from clients to data targets

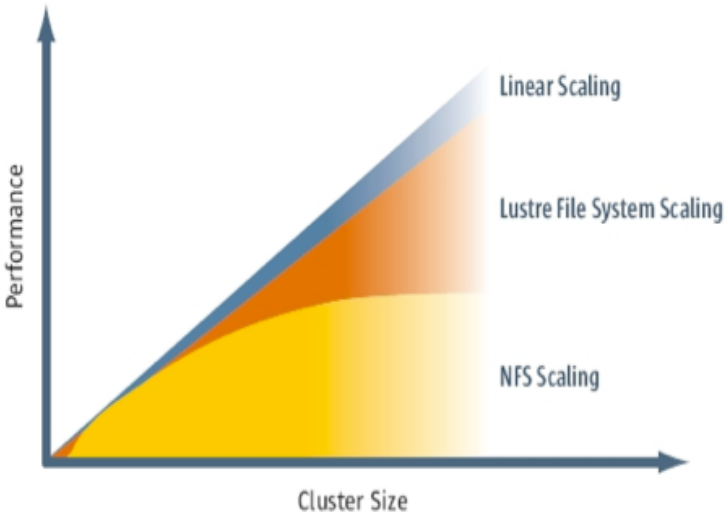
- **Bottleneck in storage server:**
  - Lustre parallel access to multiple data target
- **No easy grow.**
  - More space → adds more disks
- **SAN: HBA for clients and server**
  - Use TCP network
- **Concurrence access files: locks**
  - Control node block concurrency



1 Motivation

ÍNDICE

- **Why?**
  - **No single point of contention on IO path, data is distributed across many storage servers.**
    - 200 HDD of 1Tb are not 200Tb --> island effect
    - With lustre 200HDD of 1Tb = 200Tb
  - **Lustre is a POSIX-compliant global.**
  - **Distributed parallel filesystem.**
    - more performance access files. One file in multiple HDD.
  - **Scalable and live expandable.**
    - Append HDD to expand filesystem.
  - **Lustre is licensed under GPL.**
    - Free!!! (at least for now).



4 Availability, Reliability and Design Concerns

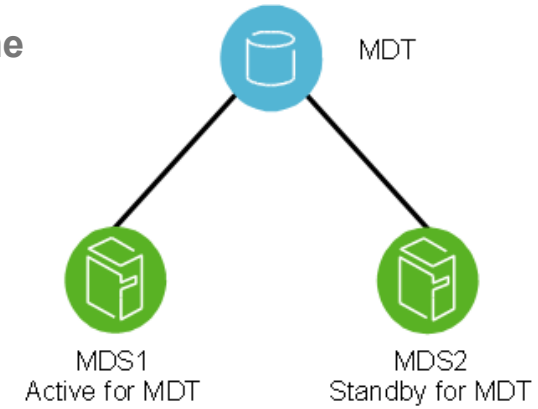
ÍNDICE

- **Availability and reliability**
  - What happens if a MDS server fail?
  - What happens if a OSS server fail?
  - What happens if a MDT server fail?
  - What happens if a OST server fail?

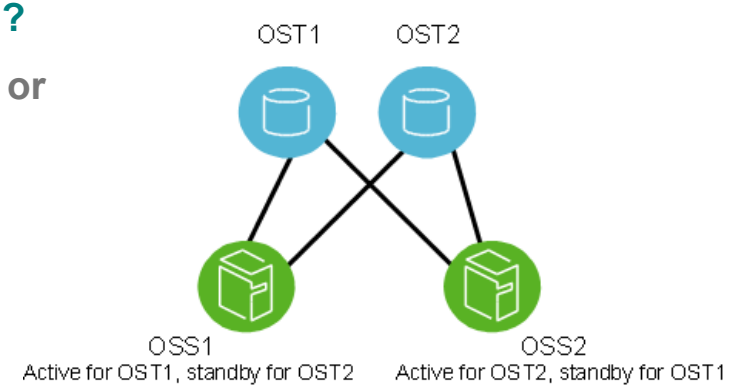
4 Availability, Reliability and Design Concerns

ÍNDICE

- **What happens if a MDS server fail?**
  - Second server in active/passive.
  - Heartbeat: MDT mounted in only one MDS.
  - STONITH, “There can be only one”.
  - Clients sense of MDS error. List of possible filesystem MDS.



- **What happens if an OSS server fail?**
  - Second server in active/active or active/passive.
  - Just-in build design.
  - Transparent for client.



4 Availability, Reliability and Design Concerns

ÍNDICE

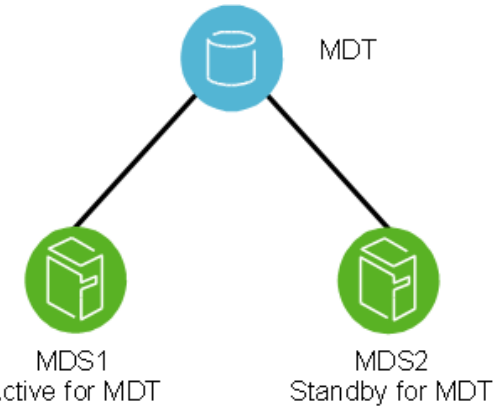
- **What happens if a MDT server fail?**
  - Single point of failure!!!!
  - Real time copy to separated device → DRBD

- **What happens if a OST server fail?**
  - Data redundancy with RAID
  - More RAID level, more reliability
  - More RAID level, less space
  - RAID failure --> lost data inside RAID

4 Availability, Reliability and Design Concerns

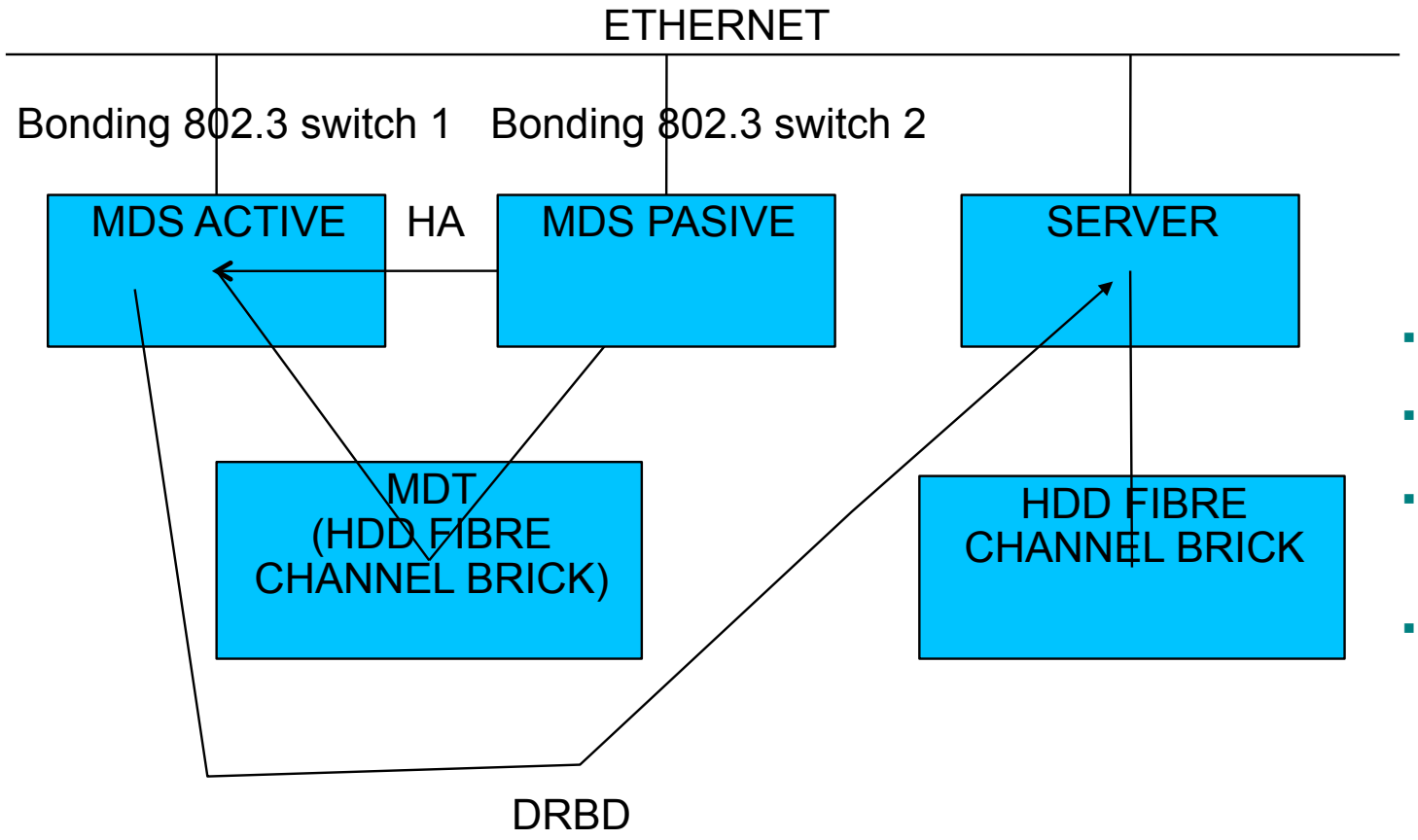
ÍNDICE

- **Design concerns: our MDS/MDT**
  - **2 IBM x336 MDS in active/passive**
    - 16Gb RAM
    - 2x3GHz CPU
    - 2GBit ethernet lacp bonding
  - **IBM DS4100 as MDT**
    - MGS integrate
    - 2Gbit fibre channel connection from MDS
    - RAID 5, one LUN per MDT
    - Hotspare for RAID



Availability, Reliability and Design Concerns

- **Design concerns: our MDS/MDT:**



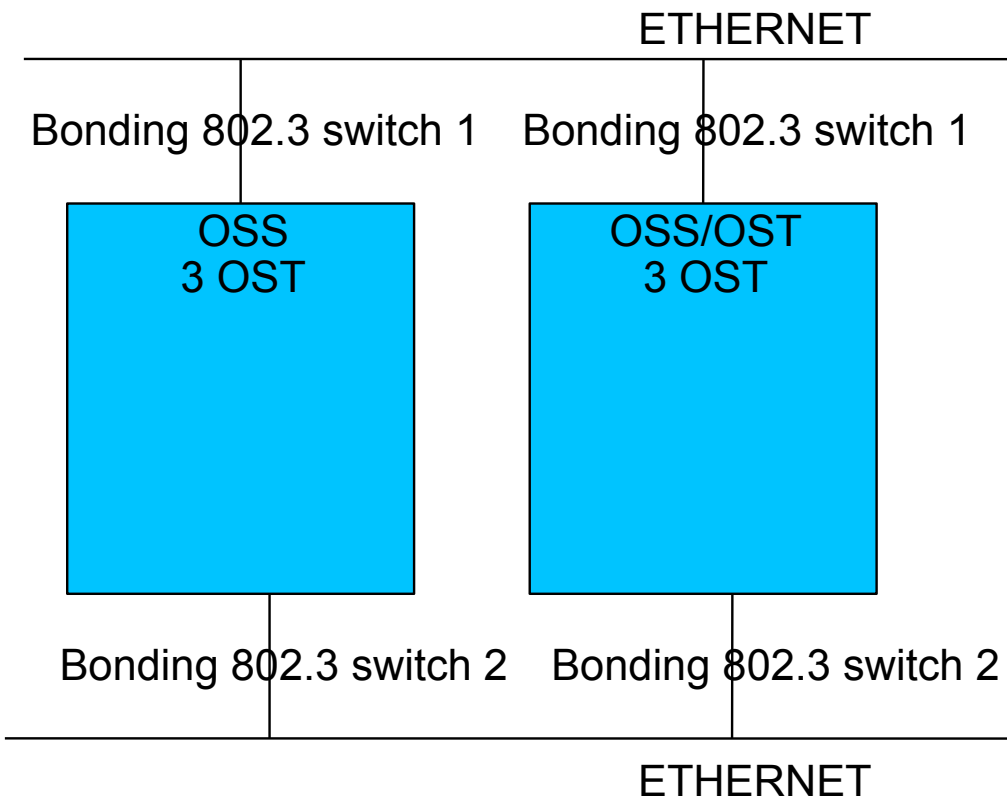
- **But…**
- **STONITH by network**
- **DRBD only MDS actived to server**
- **No load balancing**

Availability, Reliability and Design Concerns

- **Design concerns: our OSS/OST**
  - **10 Supermicro as OSS/OST**
    - 24 TB RAW => 17 TB lustre infrastructure
    - 3 RAID 5 per OST
    - HotSpare for healthy RAID
    - 8 Gb RAM
    - 2x2,5GHz CPU
    - 2GBit ethernet lacp bonding and active/passive failover bonding
  - **240 TB RAW => 171 TB lustre infrastructure**

Availability, Reliability and Design Concerns

- **Design concerns: our OSS (Supermicro):**



- **But…**
- **No OSS HA**
- **No OSS failover**
- **No OSS load balancing**

4 Availability, Reliability and Design Concerns

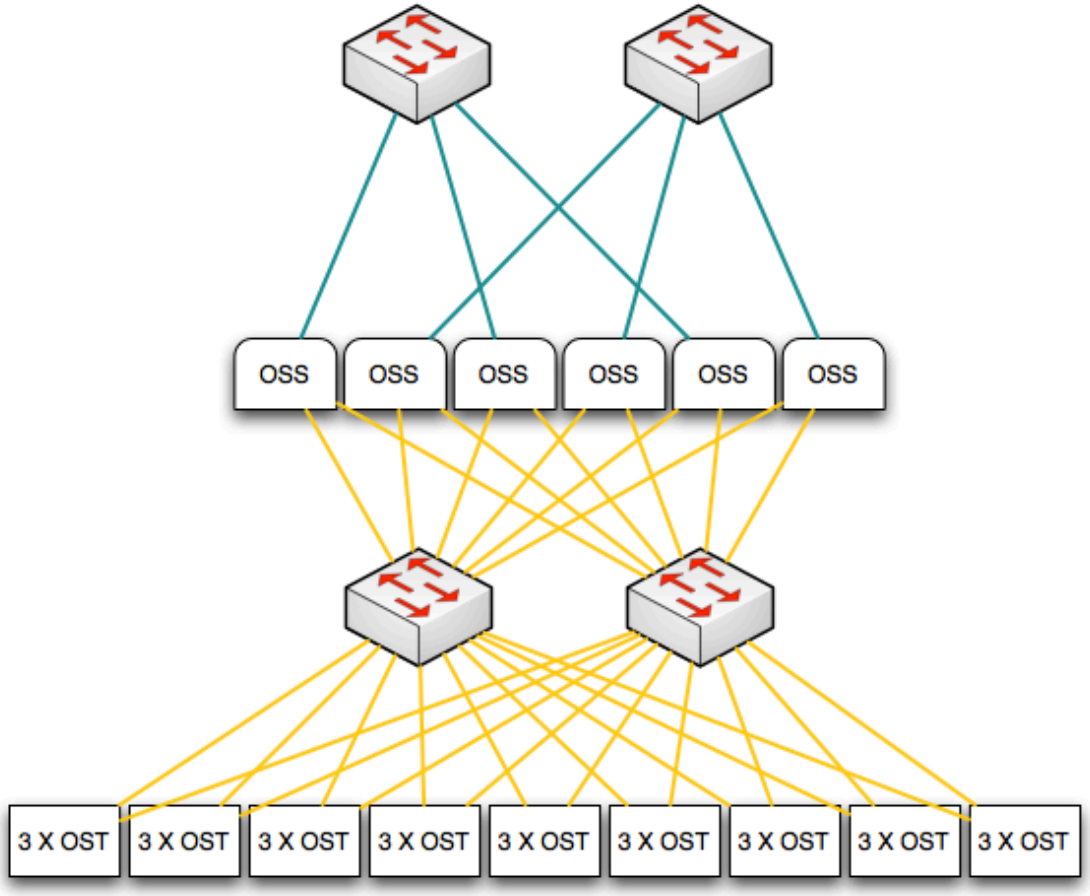
- **Design concerns: our OSS/OST**
  - **5 IBM x336 MDS in active/ passive**
    - 4Gb RAM
    - 2x3GHz CPU
    - 2GBit ethernet lacp bonding
  - **10 IBM DS4100+exp110 as OST**
    - 2Gbit fibre channel connection from OSS
    - RAID 5 per OST
    - Hotspare for RAID

ÍNDICE

4 Availability, Reliability and Design Concerns

ÍNDICE

- **Design concerns: our OSS (DS4100):**



HA:

- **Failover**
- **Load balancing → IPVS**

4 Availability, Reliability and Design Concerns

ÍNDICE

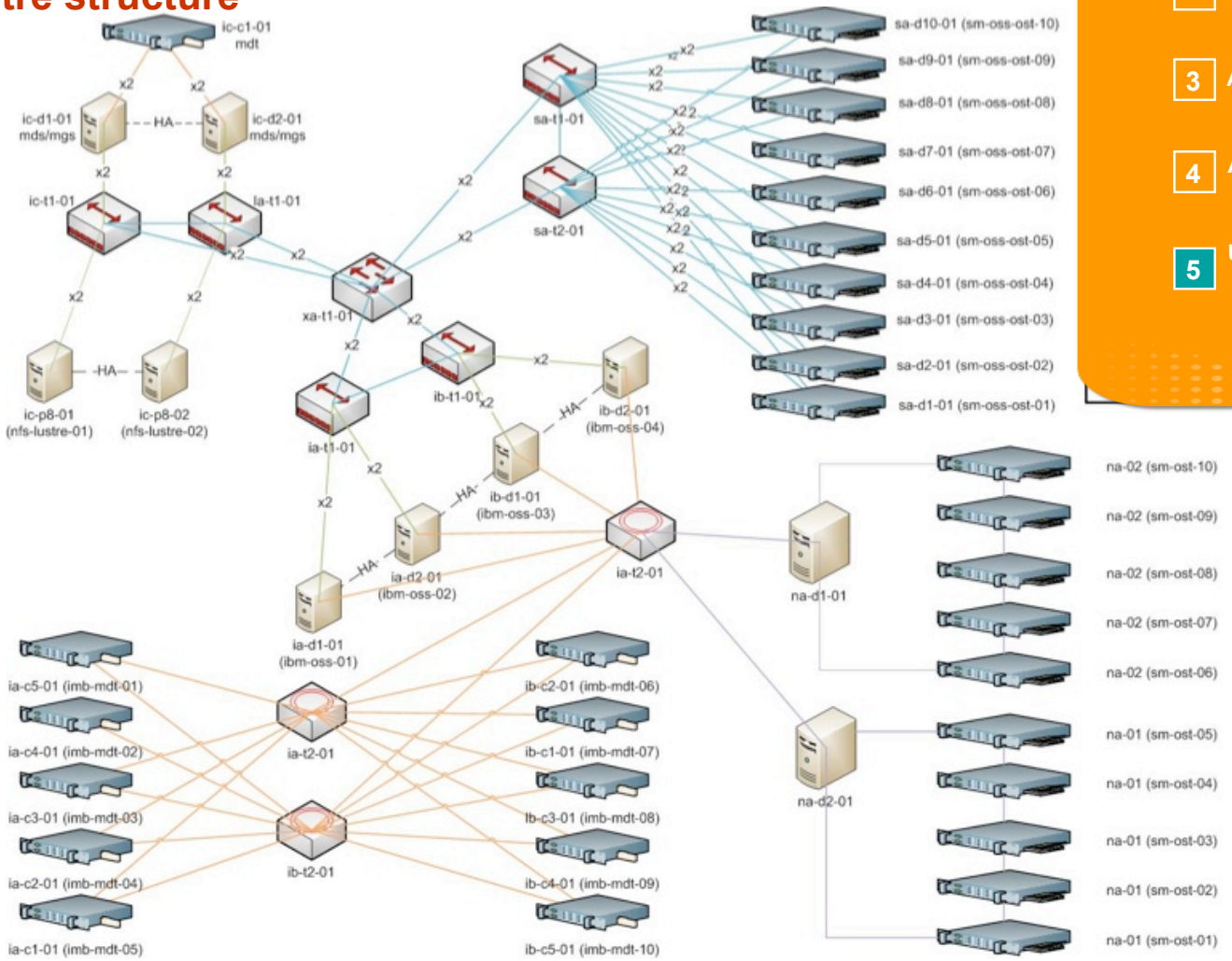
- **Design concerns: our OSS/OST**
  - **1 NetApp**
    - 168 TB RAW => 120 TB lustre infrastructure
    - 3 RAID 5 per OST
    - HotSpare for healthy RAID
    - 4Gb RAM
    - 2x2,5GHz CPU
    - IN PROGRESS…

Availability, Reliability and Design Concerns

- **Lustre version**
  - **First installation (servers and clients) → Lustre 1.8.0**
    - Poor performance
    - Critical bugs
  - **Lustre 1.8.4 (server and clients)**
    - Migration transparent (mds failover)
  - **Servers SO: Centos 5.5**
    - Recompile kernel to DRBD module support and SNMP monitoring
  - **Old machines Scientific linux 4.6 → Lustre 1.6.7.2**
    - Good integration with Lustre 1.8.4
  - **No stripe filesystem.**
    - No need speed, but needed data safe

5 Upcoming Features

- **Full lustre structure**



ÍNDICE

Availability, Reliability and Design Concerns

- **Troubles**
  - **MDS HA resource balancing → MDS Recovery mode**
    - During 600 seconds the filesystem is inaccessible --> Clients stuck
  - **DRBD is hungry of resources**
    - DRBD+MDS = 35% cpu in IDLE time
  - **Export NFS to non lustre client**
    - Needed kernel patched to export.
  - **Clients "see" the entire filesystem**

    SOLUTION:
    - "Bind mount" a filesystem subdirectory and unmount global filesystem
    - Specific order: fist mount file system, second bind mount, and at last unmount filesystem --> We can't use FSTAB
    - Specific mount script when machine starts.

Availability, Reliability and Design Concerns

- **Troubles**
  - **A cyclical restarted client can´t mount lustre:**
    - *“mount.lustre: mount mds:/filesyetem at /mnt/data failed: Cannot send after transport endpoint shutdown”*
    - Waiting for some time --> the client can mount OK
  - **Many errors but the service is 100% operative:**
    - /var/log/messages

Availability, Reliability and Design Concerns

- **Troubles**

Jul 19 13:17:53 lustre-client kernel: Lustre: 2737:0:(client.c:1476:ptlrpc_expire_one_request()) Skipped 61 previous similar messages
Jul 19 13:20:23 lustre-client kernel: Lustre: 2738:0:(import.c:517:import_select_connection()) lustre-MDT0000-mdc-df830c00: tried all connections, increasing latency to 25s
Jul 19 13:20:23 lustre-client kernel: Lustre: 2738:0:(import.c:517:import_select_connection()) Skipped 9 previous similar messages
Jul 19 13:23:03 lustre-client kernel: LustreError: 18516:0:(mdc_locks.c:646:mdc_enqueue()) ldlm_cli_enqueue: -4
Jul 19 13:23:03 lustre-client kernel: LustreError: 18516:0:(file.c:3280:ll_inode_revalidate_fini()) failure -4 inode 194478084
Jul 19 13:27:56 lustre-client kernel: Lustre: 2737:0:(client.c:1476:ptlrpc_expire_one_request()) @@@ Request x1374108762197831 sent from MGC192.168.11.9@tcp to NID 192.168.11.12@tcp 0s ago has failed due to network error (13s prior to deadline).
Jul 19 13:27:56 lustre-client kernel: req@ca00d400 x1374108762197831/t0 o250->MGS@192.168.11.12@tcp:26/25 lens 368/584 e 0 to 1 dl 1311074889 ref 1 fl Rpc:N/0/0 rc 0/0
Jul 19 13:27:56 lustre-client kernel: Lustre: 2737:0:(client.c:1476:ptlrpc_expire_one_request()) Skipped 58 previous similar messages
Jul 19 13:30:31 lustre-client kernel: Lustre: 2738:0:(import.c:517:import_select_connection()) lustre-MDT0000-mdc-df830c00: tried all connections, increasing latency to 25s
Jul 19 13:30:31 lustre-client kernel: Lustre: 2738:0:(import.c:517:import_select_connection()) Skipped 7 previous similar messages
Jul 19 13:32:36 lustre-client kernel: Lustre: 2737:0:(import.c:855:ptlrpc_connect_interpret()) MGS@192.168.11.12@tcp changed server handle from 0x8893a4118cbffab9 to 0x8893a41829f495da
Jul 19 13:32:36 lustre-client kernel: Lustre: MGC192.168.11.9@tcp: Reactivating import
Jul 19 13:32:36 lustre-client kernel: Lustre: MGC192.168.11.9@tcp: Connection restored to service MGS using nid 192.168.11.12@tcp.
Jul 19 13:33:03 lustre-client kernel: LustreError: 167-0: This client was evicted by lustre-MDT0000; in progress operations using this service will fail.
Jul 19 13:33:03 lustre-client kernel: LustreError: 18685:0:(mdc_locks.c:646:mdc_enqueue()) ldlm_cli_enqueue: -4
Jul 19 13:33:03 lustre-client kernel: LustreError: 18685:0:(file.c:3280:ll_inode_revalidate_fini()) failure -4 inode 194478084
Jul 19 13:33:03 lustre-client kernel: Lustre: lustre-MDT0000-mdc-df830c00: Connection restored to service lustre-MDT0000 using nid 192.168.11.12@tcp.

Availability, Reliability and Design Concerns

- **Mount script**
  - **/etc/fslustre:**
    - First line, main fielsystem:
      - » mds:mds_fail_over mount_point lustre options
    - Second, and later lines:
      - » lustre_subdirectory mount_point bind options
  - **mount_lustre.sh:**

```
#!/bin/bash
paso=1
while read line
do
        if [ $paso = 1 ]; then
                mds_fs=$(echo "$line" | cut -f1)
                mount_fs=$(echo "$line" | cut -f2)
                params_fs=$(echo "$line" | cut -f4)
                paso=2
                mount -t lustre -o $params_fs $mds_fs $mount_fs
        else                    dir_fs=$(echo "$line" | cut -f1)
                local_fs=$(echo "$line" | cut -f2)
                mount -o bind $dir_fs $local_fs
        fi
done < /etc/fslustreumount
$mount_fstouch /var/lock/subsys/umount_lustre
```

Availability, Reliability and Design Concerns

- **Links to start with machine:**
  - /etc/rc3.d/S60mount_lustre -> /usr/local/sbin/mount_lustre.sh
  - /etc/rc3.d/K60mount_lustre -> /usr/local/sbin/umount_lustre.sh

Troubles

- **Lustre Hardware non-commercial**
  - **Different OSS/OST with different characteristics:**
    - IBM DS4100: OSS Failover, fibre channel disk access, 2Gb RAM per OSS
    - Supermicro: Non OSS Failover, SATA disk, 4Gb RAM per OSS
    - NetApp: OSS Failover, fibre channel disk access, integrate solution
  - **Different network connection:**
    - Bonding 802.3ad or Bonding active-passive backup
  - **Different RAID controller:**
    - PERC, INTEL,… with different RAID models: RAID5, RAID DP, …

DIFFERENT PERFORMANCE

Troubles

- **Grid middleware integration**

What is grid middleware? (gLite)

- It integrates all the heterogeneous resources that are spread across multiple administrative domains (educational institutions, offices, industries) across the world.
- It gives transparent/collaborative access to these resources and has distributed Job management system.
- Grid is not a cluster: Grid grows and shrinks dynamically.

Architectural of gLite

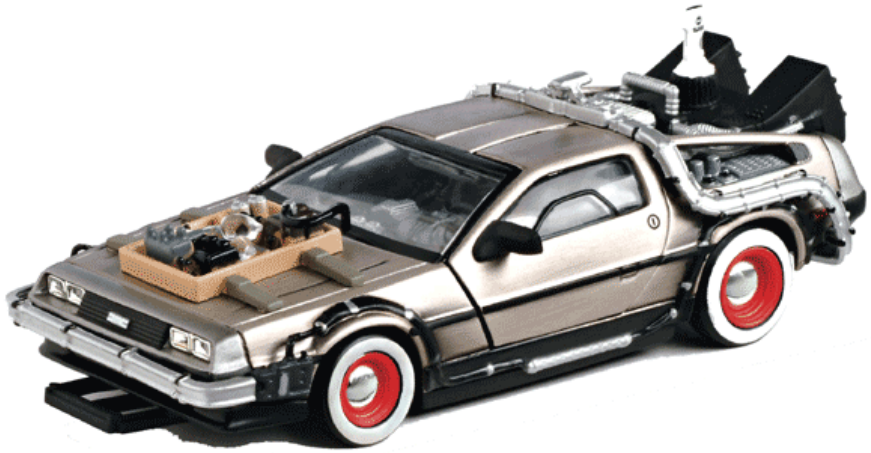
- Computing nodes: A service called “computing element” submits jobs to a worker nodes.
- Storage nodes: An user can storage data in services called “Storage nodes”.
  - » **Storage nodes manage users permissions and quotas by themselves and save data with its own user. → Trouble with ACL and QUOTAS LUSTRE**
  - » **Middleware access directly to free size storage (non posix) --> Conflict with lustre quotas and reported size to users.**

Troubles

- **Tape integration or HSF**
  - **Lustre doesn't support HSF for long life data.**
  - **We must see lustre filesystem as a big hard disk and use other software to move data.**
  - **How do we deploy Lustre client and configuration over infrastructure?**
    - **SOLUTION: Metapacket with Lustre pathless client and the client mount script.**

5 Upcoming Features

- Tape robot --> HFS (Hierarchical file system)
- Good manage: Centralized ACL, quotas, stadistics,…
- Lustre for cloud storage
- New versions: good features without bugs?



ÍNDICE

Acknowledgments

- **CETA-CIEMAT acknowledges the support of the European Regional Development Fund (ERDF/FEDER)**
  - **http://ec.europa.eu/regional_policy/funds/feder/**

ÍNDICE

Conventual de San Francisco, Sola 1, 10200 Trujillo
Phone: 927 65 93 17 Fax: 927 32 32 37
http://www.ceta-ciemat.es

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
CENTRO EXTREMEÑO DE TECNOLOGÍAS AVANZADAS
CETA Ciemat
FEDER
Fondo Europeo de Desarrollo Regional
Una manera de hacer Europa